# Shared and not shared: Providing repository services on a national level

Jyrki Ilva
National Library of Finland
e-mail: jyrki.ilva@helsinki.fi

## 1. Introduction

In March 2012 there were as many as 48 Finnish organizations with an institutional repository for their publications. These included universities, universities of applied sciences, state research institutes and scholarly societies. However, there were only twelve repository instances.[1]

Although this discrepancy may seem odd at the first sight, there is a natural explanation: most of the Finnish organizations are using shared or hosted repository services. This paper examines the reasons behind this situation, which is somewhat different from most of the other countries. It also provides some insight into what it takes to run successful shared repository services on a national level.

The National Library of Finland currently hosts two DSpace-based multi-institutional repositories, Doria (http://www.doria.fi) and Theseus (http://publications.theseus.fi), for 35 customer organizations. In addition, some of the big universities have their own repositories, most of which are also based on DSpace.

## 2. A short history of national DSpace services in Finland

The development of open repositories started out in Finland in the late 1990s. The actual term or concept had not been popularized yet, but even at this early stage there were strong incentives for disseminating scholarly publications on the Web. According to Finnish tradition, all doctoral dissertations were published as a printed book, and since the number of doctoral degrees awarded each year was growing rapidly, an alternative distribution model was needed. Although there were plans for national co-operation, several universities started their own web publishing services using various, usually non-ambitious technical solutions. The National Library was involved in this development by running the E-thesis service for the University of Helsinki.



In 2003 the National Library came up with a new plan for the national architecture of future digital library systems. One of the needs identified in this plan was a digital object management system which would work seamlessly together with library catalogs and portals. The first attempt to address this need was made with a proprietary software platform chosen by competitive tender. Unfortunately it filled the requirements only on paper, and in practice didn't work out as has been planned.

To salvage the situation the National Library was forced to look for other solutions. Open source repository software platforms had improved quite a bit in a few years, and in April 2006 the National Library made a short evaluation of DSpace, EPrints, Fedora and CDSware. Quick implementation and suitability for multi-institutional use were considered central selling points, and thus DSpace was chosen.

The first DSpace instance, Doria, went live with first collections in February 2007. The idea was to collect all of the stuff into one big DSpace instance, which was thought to be easier to manage and more convenient for end users than many separate instances. To facilitate the multi-institutional use the National Library was one of the early adopters of Manakin, which made it possible to modify the look of communities and collections. Another issue that required special attention was the multi-lingual (Finnish, Swedish and English) user interface.

With many important national projects competing for attention, the DSpace project was not one of the top priorities at the National Library. It received only relatively modest, temporary project funding from the Ministry of Culture and Education. Without stable centralized funding it was obvious that most of the money would have

[1] The number includes three separate instances run by the University of Tampere.

to come from the participating organizations. However, the customer-base of the new service was at first still quite small, about ten organizations of various sizes, which was too little to support the kind of service that the National Library was trying to provide.

The lucky break came in late 2007 when the Rector's Council for Universities of Applied Sciences got the funding for Theseus, the common repository for all of the 25 universities of applied sciences. The National Library was chosen to be the service provider, and the work started in the early months of 2008. The scale of the project was quite ambitious, but with some effort all of the practical details were successfully ironed out so that the new service could be adopted in all 25 organizations by 2010.

Another lucky coincidence was the *Cultural Materials Depositing and Preservation Act* (2008), which gave the National Library new duties in web archiving, and also led to the purchase of a new server infrastructure. This benefited the repository services as well, as there was a chance to rebuild the whole system in a new, more robust and much more flexible virtual server environment. The adoption of SVN version control in 2010 made it much easier to maintain many parallel DSpace instances with the same standardized code base. One consequence of this was that it was now easy to offer the customers the possibility to have their own hosted DSpace instance for only a small extra cost.

## 3. Pros and cons of shared repository services

The commonly accepted view of institutional repositories includes an assumption that repositories are built on an organizational level and the content of these local repositories is harvested via OAI-PMH to centralized search engines specializing on scholarly content. In practice a large majority of repositories have been based on open source software, and the whole approach has generally been very much DIY-oriented: anyone can set up a repository instance and every organization should have one. As a result we have a large global network of mostly separately-hosted repository software instances, many of which have relatively little content in them.

It has been proven that anyone with a bit of technical know-how and resources can set up a working repository instance, but does this really make sense? It is obvious that there is a lot of duplication of effort in running all of these separate instances, and it is possible to argue that some of these resources would be better spent elsewhere. Another issue to consider is the viability of small repository instances, many of which are dependent on the expertise of one technical person, who may at some point decide to move on to other things. In many cases it might be worth considering whether the success of an institutional repository service really requires the use of a locally-run repository instance, or whether it would make more sense to outsource the maintenance of the technical platform or join a larger consortium to maintain one.



These arguments clearly appeal to funding bodies and to some extent also to library leadership, but it has been much harder to persuade repository managers - especially those with technical ambition - to agree. Some of the reasons may be psychological: one of the relatively few rewards in being a "repository rat" (to borrow a phrase from Dorothea Salo) is the opportunity to do things independently. There may even be an amount of romanticism associated with this view. Outsourcing some of the work to a consortium or an outside service-provider sounds much less heroic. Depending on the repository rat's position at the organization it may also sound dangerous: "if all of this work is outsourced, do they still need me"?

Of course, there are many valid concerns associated with the use centralized services. Is the service provider trustworthy? Are the services flexible enough to adapt to my organization's specific needs? Does the use of a hosted service support the visibility of our university brand? If everything is centralized, is there still room for innovation? However, these are all concerns that can be taken into account in the planning of the service, and in the division of work between the local repository managers and the service provider or consortium.

## 4. The case of shared repository services in Finland

The idea of building shared repository services seems to have been much more successful in Finland than in most of the other countries. This is no coincidence. Finland is a relatively small country with a uniform higher education sector. Finnish libraries are also used to working together and building shared services.

The National Library of Finland is a central service provider for the libraries: it takes care of the management of university library catalogs, portals and national consortia for the acquisition of electronic books and journals

(FinELib). The role of the Ministry of Culture and Education has been an important reason for this: the Ministry is the main funder of nearly all Finnish universities, and consequently it has had a strong incentive to support the creation of efficient and cost-effective national services which benefit the whole higher education sector.

The development of a national repository infrastructure has not been quite straightforward, though. Several of the Finnish universities had been building their own repository services before the National Library started to offer centralized services, and most of these universities have chosen to continue with their own services, which are in some cases very sophisticated. Although the use of the infrastructure provided by the National Library would be in most cases considerably cheaper, the repository managers and library leadership have judged that having the whole process within their own organization provides added value that justifies the extra cost.

It seems likely that some of the universities are going to have their own locally-run repositories even in the future. This doesn't rule out the need for co-operation between the repositories and repository managers. The National Library has been promoting national co-operation by organizing meetings, seminars and other venues for the exchange of ideas. The main problem slowing down the co-operation seems to be lack of resources on the local level. Although there is a lot of interest in sharing ideas, code and best practices, many repository managers have a large amount of daily work (some of which is not related to repositories) in their hands, and to them co-operation with other repository managers may seem to be a luxury that they cannot afford.

## 5. Three ways of providing hosted repository services

The service model offered by the National Library to its repository customers is based on division of work. Most of the work is done locally in each customer organization, as they take care of the curation of their own collections and publications. On the other hand, the National Library is responsible for the development and maintenance of the technical platform. In addition to the basic repository services, the National Library can upon request also provide consulting and build extra services to meet more specialized customer needs.

The customer organizations may either use one of the multi-institutional repository instances (Doria, Theseus) or their own DSpace instance hosted by the National Library. Although the technical maintenance of the repository instances is highly centralized, there is currently quite a bit of variety in the level of standardization of processes between the repository instances. Three different models can be identified:



1) Theseus: a multi-institutional repository instance with standardized processes

> All of the 25 organizations participating in Theseus use the same tools, formats and processes, and all publications - more than 10.000 a year - are saved into one DSpace instance. In essence this is a large group effort: there are more than two hundred librarians and administrators in the participating organizations handling the new theses that are submitted by the students. All collections have the same uniform appearance, and the standardization of processes and interfaces decreases the maintenance cost of the service significantly.
>
> The centralized resources needed to make Theseus work are fairly small. There is a small team of people from a handful of libraries who can devote some of their working time to managing and co-ordinating the service on the national level. The technical maintenance of the service obviously benefits from the work that the National Library is doing for all of its DSpace instances, but it needs relatively little dedicated resources (less than one man-year, actually) for the technical maintenance and development. Compared to how much it would cost to manage 25 repository instances separately this is an extremely cost-effective way of doing things.

2) Doria: a multi-institutional repository instance with a variety of processes

> Doria was originally designed to be a neutral technical platform which any organization could easily adopt for its own use. The name and URL of the service were chosen so that they were not associated with the National Library or any other organization. Some of the collections and processes were transported from other pre-existing systems, and the customer organizations were given relatively free hands in using their own visual themes and metadata formats. Some of the organizations use the built-in input forms of DSpace, some of them have their own external submission processes.

The downside of the lack of standardization is that the quality of the item metadata in Doria is not uniform, and the customized community-level user interfaces may appear confusing to end users.

3) Separate repository instances

With the new virtual server infrastructure and SVN version control the National Library is now in a good position to host separate repository instances for individual customer organizations. Many of the bigger organizations with a large number of publications prefer this solution, as it seems to give greater visibility to their own brand. In some cases separate instances may also be easier to maintain, especially if there is a lot of customization, and the repository is connected to a large number of external information systems.

The first instance of this kind is the institutional repository of Finland's National Institute for Health and Welfare, which may later grow to include other organizations funded by the Ministry of Social Affairs and Health. There are also two other separate repositories in the early stages of planning: one of them will be built for a state research institute, and one for a university.

The National Library has been working on a new external ingest system (SYLI), which will make it easier to modify the ingest processes to suit the needs of various customer organizations. The ingest system will be connected to DSpace via a REST interface, and it is currently in test use. It is hoped that the new ingest system will also help in improving the quality of metadata.

## 6. Measures of success

The public discourse on open repositories has usually concentrated on the self-archiving of scientific publications. This discourse has been very successful in boosting the credibility of repositories and in many cases it has provided the main motivation for the setting up of a repository. However, the number of articles that the scholars have actually self-archived has remained relatively low, and it has not been the central usage case for most of the repositories. In Finland there are currently four institutional self-archiving mandates in effect, but the total number of self-archived articles is still fairly small.



On the other hand, the open access publication of theses and dissertations has become a success story. The Finnish repositories currently contain almost 70.000 open access theses, which is about two thirds of all full-text content stored in these repositories. Most of the new doctoral dissertations are now open access publications, and the amount of other theses published on the Web is growing as well.

It's useful to remember that having a repository is not an end to itself; the repository is a tool that can be used to achieve various objectives. Although self-archiving has so far achieved only modest success, and the shortcomings of repository software have also received their share of criticism, there are still valid reasons to believe in the viability of the open repository concept. Our experience in working with customer organizations shows that there is a strong practical need for an affordable system that can be used for the storing and dissemination of digital materials. These may be theses, serial publications, cultural heritage materials, research data or self-archived articles.

Although current repository software platforms may not be ideal choices for all purposes, they are still good enough in most cases. Our customer organizations generally expect that the repository service can provide reliable long-term access to their publications, preferably with persistent addresses. It is also important that the repository service can be easily integrated with other information systems. In many Finnish universities and state research institutes integration with current research information systems (CRIS) has been one of the key requirements for a repository service. Another common requirement is the possibility to harvest metadata from other systems to make the ingest process easier.

There are different measures for the success of a repository service. Doria and Theseus currently contain more than 55.000 open access full-text items of various categories, and according to the annual 2011 statistics, full-text files stored in these two repository instances were downloaded more than eight million times. However, it is not quite as easy to provide exact figures for the value of these services for participating organizations, or for their value in increasing the impact of the materials stored in them.